Old Dominion University
ODU Digital Commons

Computer Science Presentations | Computer Science

5-21-2014

Profiling Web Archives

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

Ahmed Alsum
Old Dominion University

Michele C. Weigle
Old Dominion University, mweigle@odu.edu

Herbert Van de Sompel

David Rosenthal

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_presentations

 Part of the Archival Science Commons

Recommended Citation
Nelson, Michael L.; Alsum, Ahmed; Weigle, Michele C.; de Sompel, Herbert Van; and Rosenthal, David, "Profiling Web Archives" (2014). *Computer Science Presentations*. 9.
https://digitalcommons.odu.edu/computerscience_presentations/9

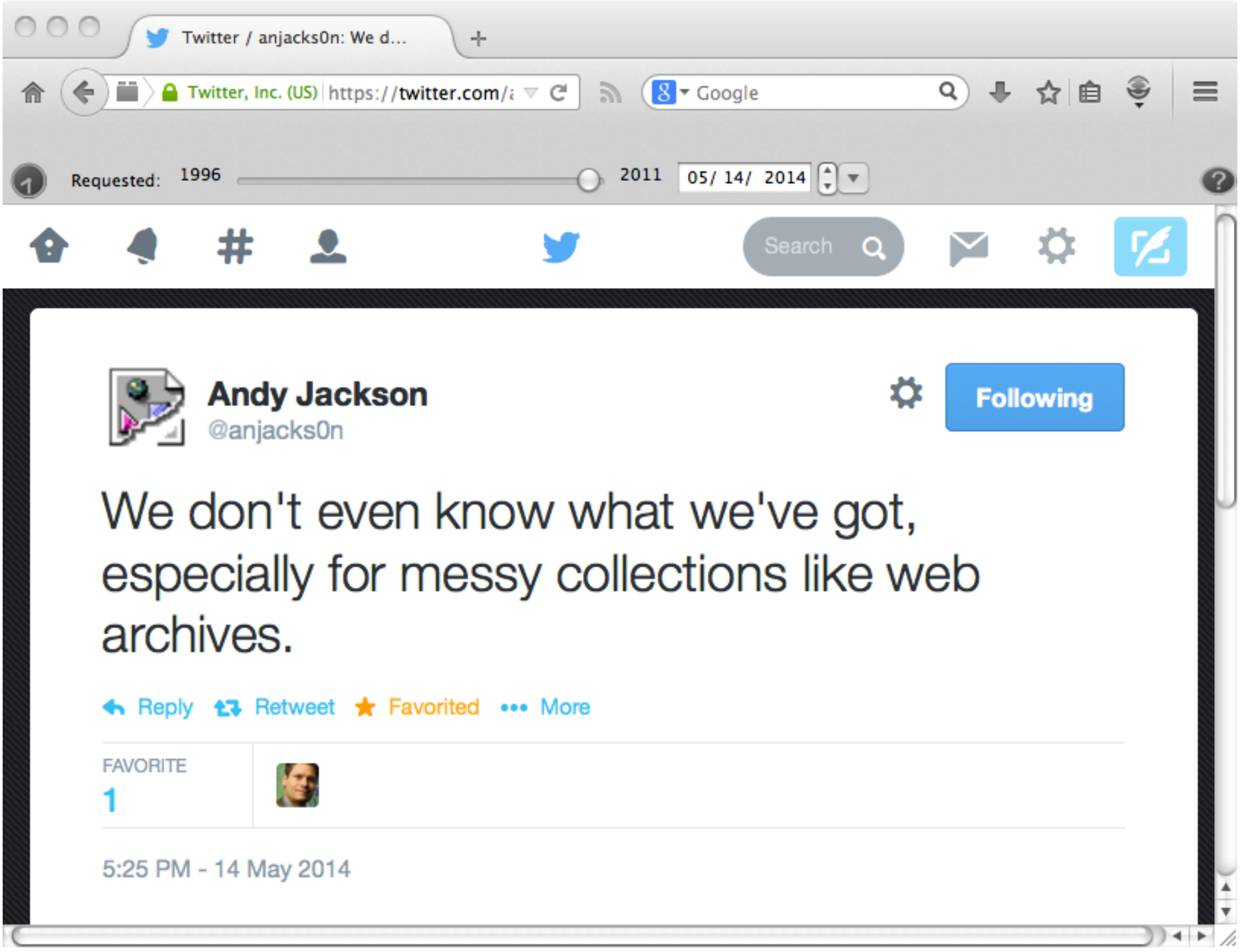
This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

Profiling Web Archives

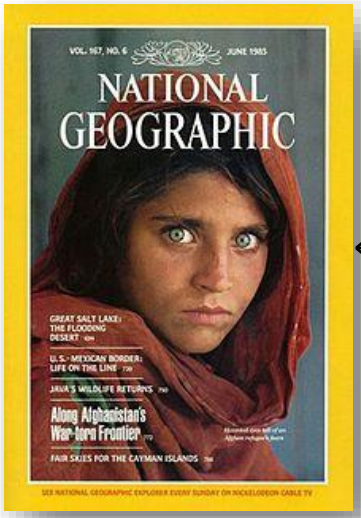
Michael L. Nelson
Ahmed AlSum, Michele C. Weigle
Herbert Van de Sompel, David Rosenthal

IIPC General Assembly
Paris, France, May 21, 2014

CC BY NC SA


Twitter / anjacks0n: We d...
Twitter, Inc. (US) https://twitter.com/a
Google
Requested: 1996 2011 05/ 14/ 2014
Search
Andy Jackson
@anjacks0n
Following
We don't even know what we've got, especially for messy collections like web archives.
Reply Retweet Favorited More
FAVORITE
1
5:25 PM - 14 May 2014


CUTTY SARK
SCOTCH WHISKY
SHARP



Where's that issue with the Afghan girl?


LIBRARY OF
CONGRESS


INTERNET ARCHIVE


Czech Web
WebArchiv


臺灣大學網站典藏庫
NTU Web Archiving System


PORTUGUESE
WEB ARCHIVE


NLB
National Library
Singapore


ARCHIVE-IT

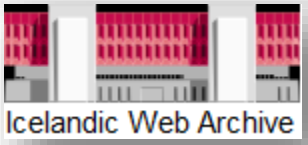

BRITISH
LIBRARY


Archive.is


WebCite


archiefweb.eu


The National Archives


Icelandic Web Archive


Library and Archives
Canada


padicat
The Web Archive of Catalonia


spletni
ARHIV
NUK


HAW
Hrvatski arhiv weba
Croatian Web Archive


LIBRARY OF
CONGRESS


INTERNET ARCHIVE


Czech Web
WebArchiv


臺灣大學網站典藏庫
NTU Web Archiving System


PORTUGUESE
WEB ARCHIVE


NLB
National Library
Singapore


memento


ARCHIVE-IT


BRITISH
LIBRARY

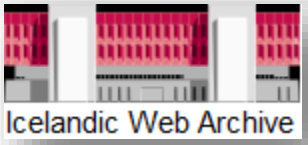

Archive.is


WebCite


archiefweb.eu


The National Archives

Aggregator

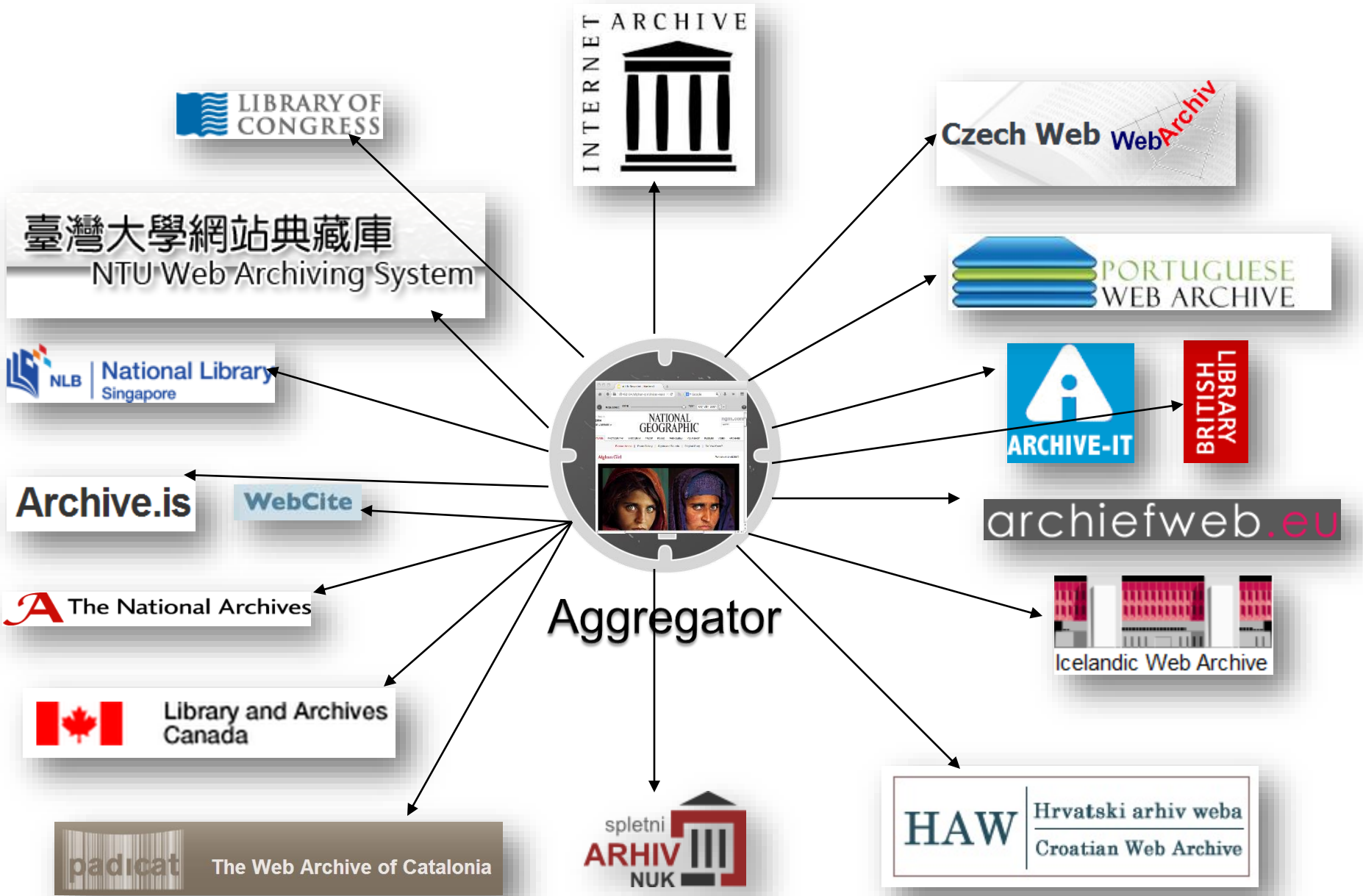

Icelandic Web Archive


Library and Archives
Canada


padicat
The Web Archive of Catalonia


spletni
ARHIV
NUK


HAW
Hrvatski arhiv weba
Croatian Web Archive


INTERNET ARCHIVE
LIBRARY OF CONGRESS
Czech Web WebArchiv
臺灣大學網站典藏庫
NTU Web Archiving System
PORTUGUESE WEB ARCHIVE
NLB National Library Singapore
ARCHIVE-IT
BRITISH LIBRARY
Archive.is
WebCite
archiefweb.eu
The National Archives
Aggregator
Icelandic Web Archive
Library and Archives Canada
spletni ARHIV NUK
HAW Hrvatski arhiv weba Croatian Web Archive
padicat The Web Archive of Catalonia

Prior IIPC Memento Aggregator Project

- Ten IIPC archives, led by LANL
- Conceived at 2011 IIPC meeting
- Results reported at 2012 IIPC meeting
  - http://netpreserve.org/sites/default/files/resources/Sanderson.pdf
- Two highlights:

The Plan…

- To provide fast access to distributed archives, LANL would merge the indexes of the holdings of multiple archives and provide Memento based access

- Step 1: Library of Congress gathers CDX files
  Step 2: LANL indexes (…)
  Step 3: Profit

- Data: 5T of gzipped CDX files (mostly from IA)
  - Shipped on hard drives
- Computing: 210 node cluster at LANL
  - 2x 2ghz processors, 2x 2T HDD, 8G RAM

… and the Reality

- Hardware failure killed one of the drives en route
  - Transferred remaining files via BagIt from LoC
- Compute cluster has very restricted access:
  - Had to transfer data over infranet
  - 2 weeks to sync (5Mb/sec)
  - And then 2 weeks to get the processed results off
- Compute cluster has faulty switch, unreliable nodes:
  - Ran original processing 15 times without success due to hardware failures

Stop and Rethink…

- LANL's processing was informative from a "big data" perspective, but was neither scalable nor sustainable
  - "send us your CDX" == hard for both parties
  - there are lots of URIs in the world
- Will only get worse with:
  - more archives…
  - …doing more archiving

Leverage Memento Aggregators

- Memento aggregator currently broadcast URI lookups to all known archives

`http://mementoproxy.lanl.gov/aggr/timemap/link/1/http://www.bnf.fr/`

- New approach:
  1. build profiles based on sampling from URI lookups (optionally supplement with CDX files when available)
  2. Use archive profiles for informing Memento aggregator "query routing" decisions
  3. *Share serialized profiles with other IIPC partners*

Profiling Studies

- TPDL 2013
  - 12 archives, March 2013, public web archives used but techniques apply generally
  - sampling only, no CDX access
- IJDL 2014 (to appear)
  - 15 archives (+4, -1), October 2013
  - slightly larger sample URI dataset
  - results similar

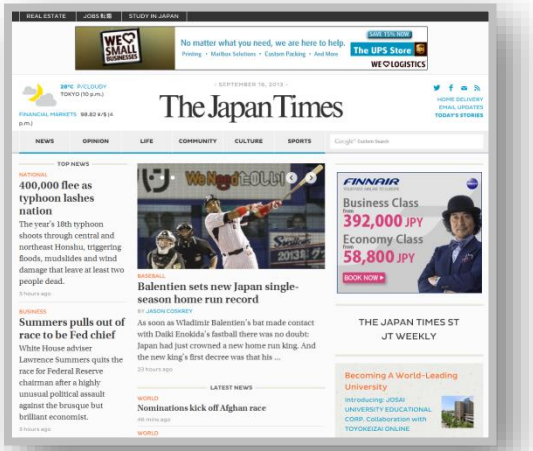
URI Lookup = Limited Information

```
GET /aggr/timegate/http://www.bnf.fr/ HTTP/1.1
Host: mementoproxy.lanl.gov
Accept-Datetime: Sun, 29 May 2005 02:46:53 GMT
Accept-Language: fr; q=1.0, en; q=0.5
…
```

1
2
3

1. Original URI
2. Memento-Datetime
3. Preferred URI

Where to find Mementos for …



http://www.japantimes.co.jp/

NLB National Library Singapore

PORTUGUESE WEB ARCHIVE

The National Archives

臺灣大學網站典藏庫
NTU Web Archiving System

INTERNET ARCHIVE

BRITISH LIBRARY

spletni ARHIV NUK

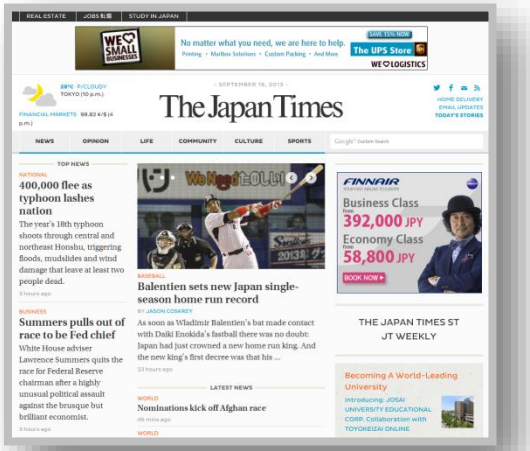
ARCHIVE-IT

Icelandic Web Archive

Czech Web WebArchiv

Archive.is

Where to find Mementos for …



http://www.japantimes.co.jp/

NLB National Library Singapore

The National Archives

Icelandic Web Archive

PORTUGUESE WEB ARCHIVE

臺灣大學網站典藏庫
NTU Web Archiving System

Czech Web WebArchiv

INTERNET ARCHIVE

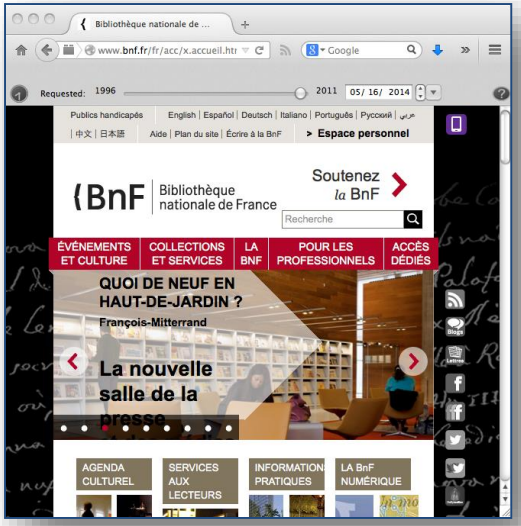
BRITISH LIBRARY

spletni ARHIV NUK

ARCHIVE-IT

Archive.is

Where to find Mementos for …



http://www.bnf.fr

NLB National Library Singapore

PORTUGUESE WEB ARCHIVE

The National Archives

臺灣大學網站典藏庫 NTU Web Archiving System

INTERNET ARCHIVE

BRITISH LIBRARY

spletni ARHIV NUK

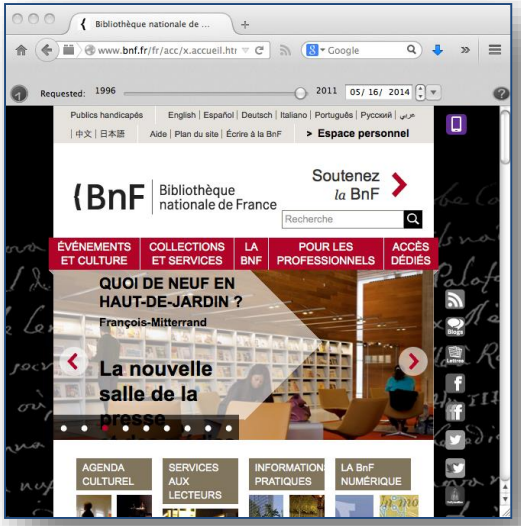
Icelandic Web Archive

Czech Web WebArchiv

ARCHIVE-IT

Archive.is

Where to find Mementos for …



http://www.bnf.fr

NLB National Library Singapore

PORTUGUESE WEB ARCHIVE

The National Archives

臺灣大學網站典藏庫 NTU Web Archiving System

INTERNET ARCHIVE

BRITISH LIBRARY

spletni ARHIV NUK

ARCHIVE-IT

Icelandic Web Archive

Czech Web WebArchiv

Archive.is

Research Question

Problem

- Profile public web archives according to the following dimensions:
  - Top-level domains
  - Languages
  - Growth rate
  - Archival date

Motivation

- Determine who is archiving what
- Optimize query routing for a Memento Aggregator

Web Archives in this Experiment

	Full text	URI-lookup
Internet Archive		√
Library of Congress		√
Icelandic Web Archive		√
Library and Archives Canada	√	√
British Library	√	√
UK National Library	√	√
Portuguese Web Archive	√	√
Web Archive of Catalonia	√	√
Croatian Web Archive	√	√
Archive of the Czech Web	√	√
National Taiwan University	√	√
Archive It	√	√

Experiment Set Up

- Sample URIs from seven different sources
- Retrieve the TimeMap for each URI from all archives
  - A TimeMap lists all Mementos for a given URI
  - A Memento is an archived version of a resource
- Analyze who has holdings for which URIs

Sampling URIs - DMOZ

1. DMOZ:Random
   - 10,000 URIs randomly sampled from DMOZ directory (~5M URIs).

2. DMOZ:TLD - 2% for each TLD from DMOZ or 100 URIs whichever is greater
   - 52 TLDs (**com** 23,470) (**de** 6,332), (**org** 4,025), (**uk** 3,309), (**net** 2,073), (**it** 1,775), (**jp** 1379), (**ru** 1244), (**fr** 1154), (**pl** 1062), (**au** 764), (**ca** 642), (**at** 438), (**edu** 390), (**cz** 385), (**tr** 334), (**info** 319), (**cn** 278), (**us** 266), (**nz** 265), (**es** 238), (**ar** 213), (**no** 150), (**br** 149), (**tw** 141), (**za** 118), (**fi** 113), (100 URIs for [**ae**, **cat**, **cl**, **cu**, **eg**, **gov**, **id**, **in**, **ir**, **is**, **ke**, **kr**, **ma**, **mt**, **mx**, **my**, **na**, **pe**, **pk**, **pt**, **sa**, **to**, **uy**, **zw**])

3. DMOZ:Languages - 100 URIs for each language
   - 24 languages: Icelandic, Portuguese, Catalan, Afrikaans, Arabic, Indonesian, Chinese (Simplified), Chinese (Traditional), Dutch, Spanish, French, Greek, Hindi, Italian, Japanese, Korean, Norwegian, Persian, Polish , Russian, Turkish, Ukrainian

Sampling URIs – Web Archives Full Text

- Query the fulltext search interface of select web archives with two sets of query terms.

4. Top 1-Gram from Bing
    - Most are English

5. Top 1000 query terms from Yahoo in 9 languages
    - Excluding general keywords such as: Obama, Facebook.

Sampling URIs – Web Archives Full Text

		Yahoo	Bing
Archive with FullText search	AIT	12617	3953
	BL	6430	3187
	CAN	1351	1107
	CR	1599	1201
	CZ	6081	3360
	CAT	8996	4241
	PO	14126	5004
	TW	1004	354
	UK	8261	3431

Sampling URIs – Web Archives Full Text

		Chinese	English	French	German	Italian	Japanese	Korean	Portuguese	Spanish	Yahoo
Archive with FullText search	AIT	26	2066	**3512**	**3837**	3321	119	2	2434	2141	**12617**
	BL	163	2354	**2350**	**2240**	**2068**	225	131	**1940**	**2056**	**6430**
	CAN	49	800	804	646	601	77	113	580	514	**1351**
	CR	54	706	697	703	701	74	19	599	600	**1599**
	CZ	**363**	1782	1578	1695	1519	**577**	114	1310	1278	**6081**
	CAT	28	2775	2496	2448	2280	209	129	2164	2429	**8996**
	PO	91	2460	3603	3081	3113	53	69	3267	3177	**14126**
	TW	357	178	176	165	157	106	7	198	119	**1004**
	UK	0	2698	**2009**	**2049**	**2046**	0	0	1903	1871	**8261**

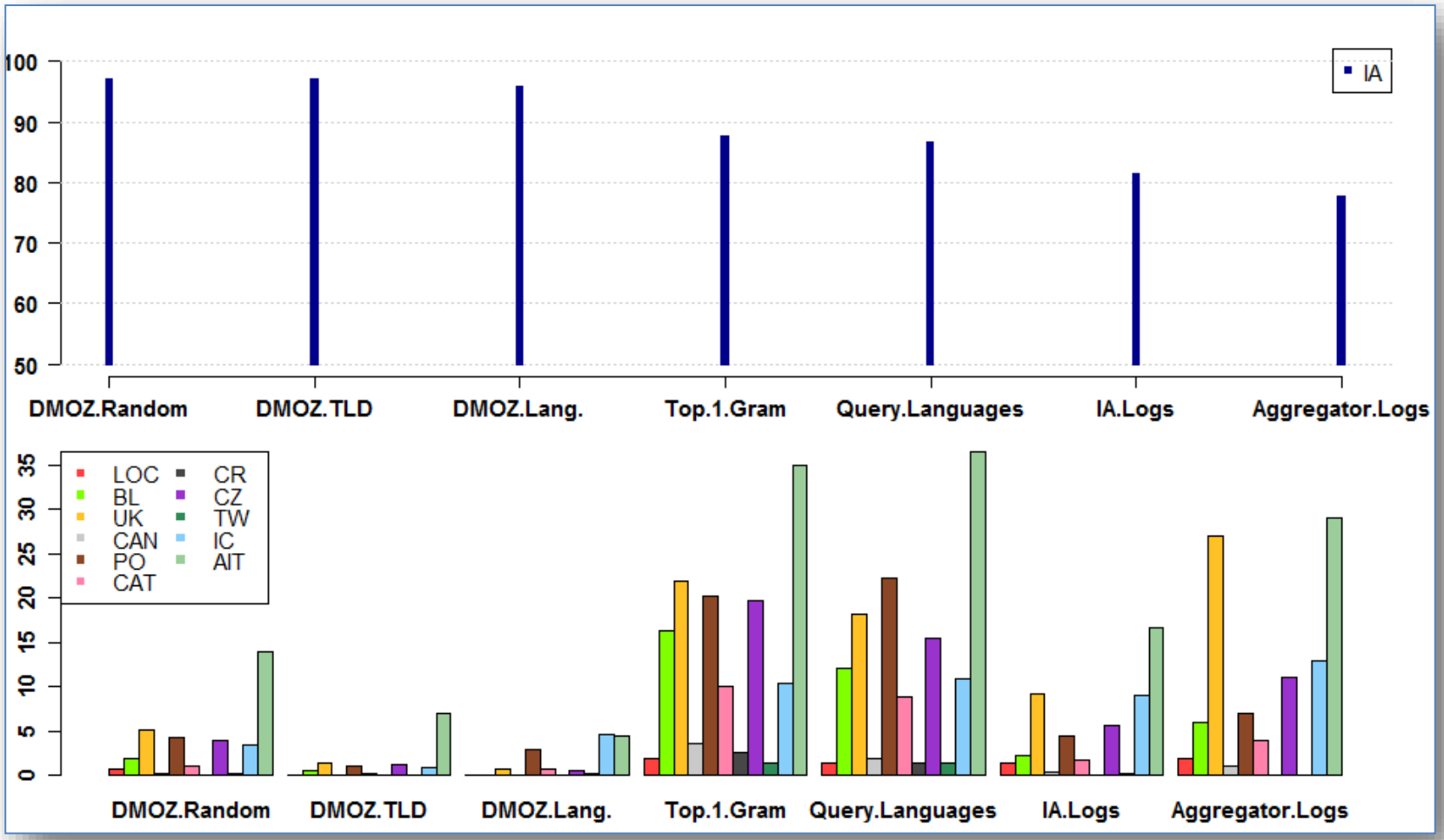
Sampling URIs – User Requests

- Sampling from user requests for archived web resources

6. Sample from IA Wayback Machine Log files
    - 1,000 URIs randomly sampled from Feb 22, 2012 to Feb 26, 2012.
7. Sample from Memento Aggregator log files
    - 100 URIs randomly sampled from LANL Memento Aggregator between 2011 to 2013.

Archive Coverage per Sample

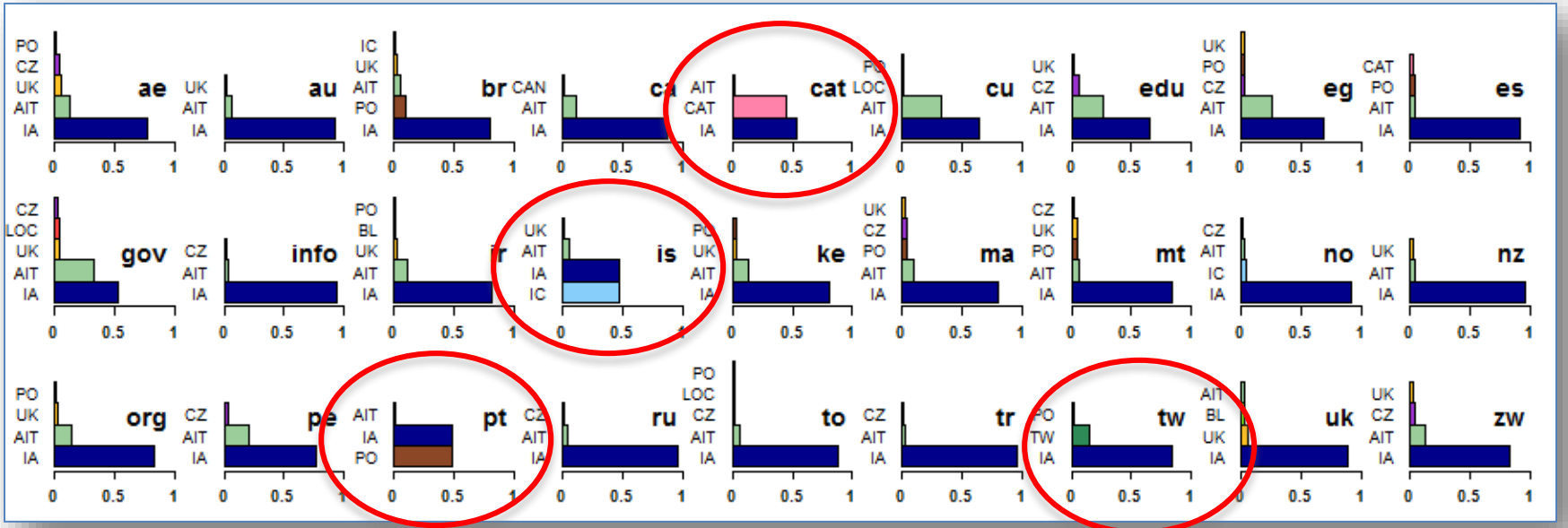
100%



35%

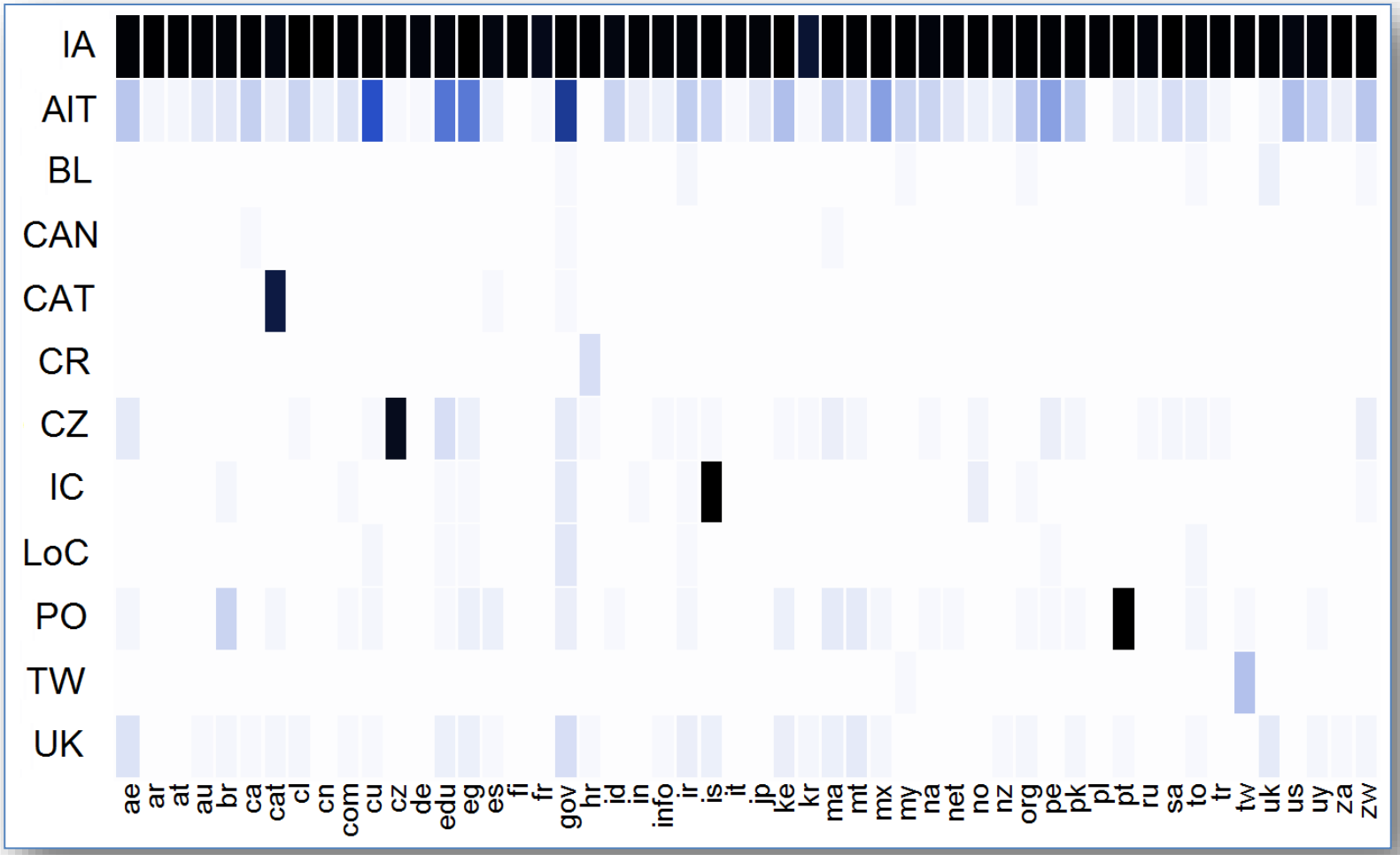
Entire Sample

TLD Coverage across Archives (1)



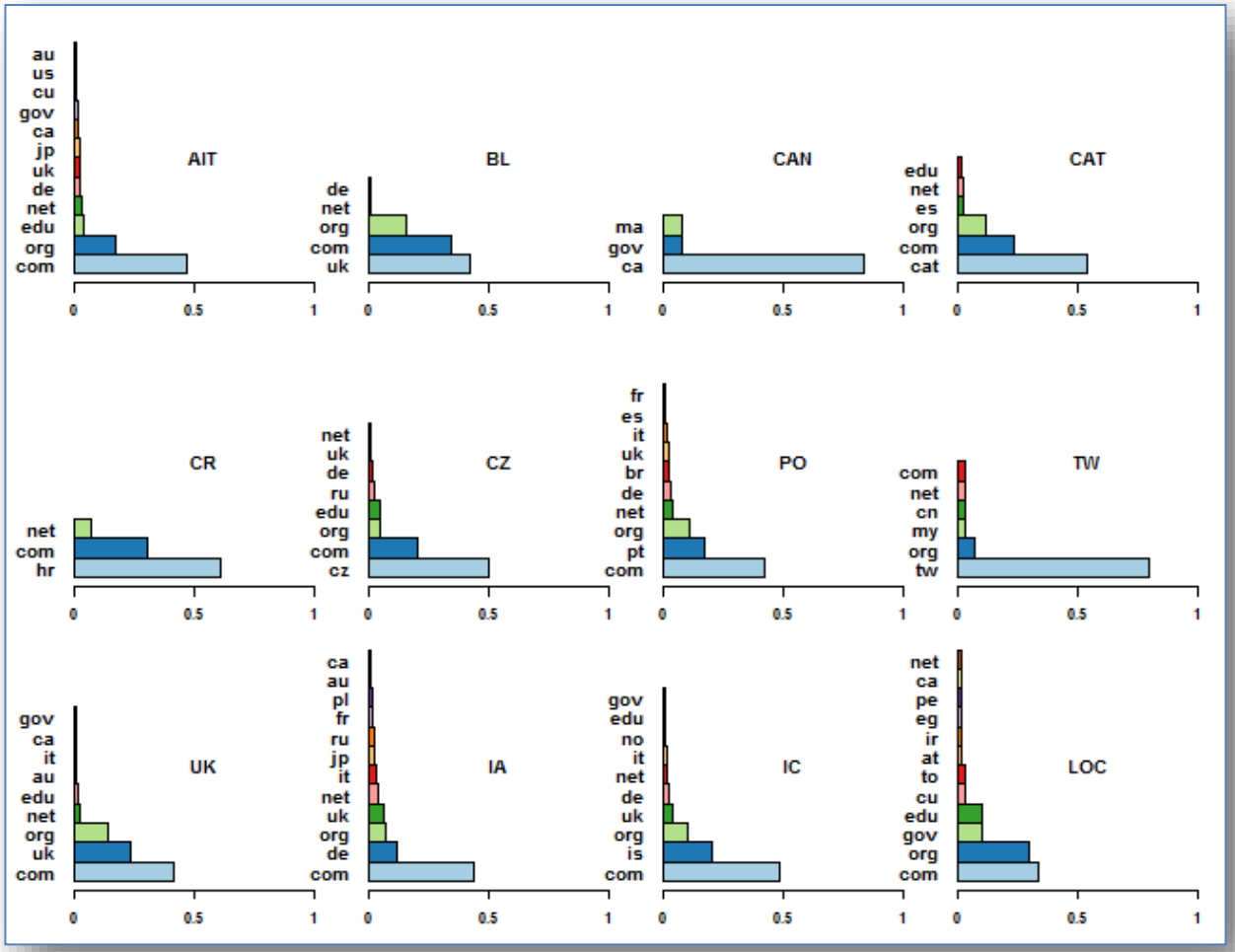
Entire Sample

TLD Coverage across Archives (2)



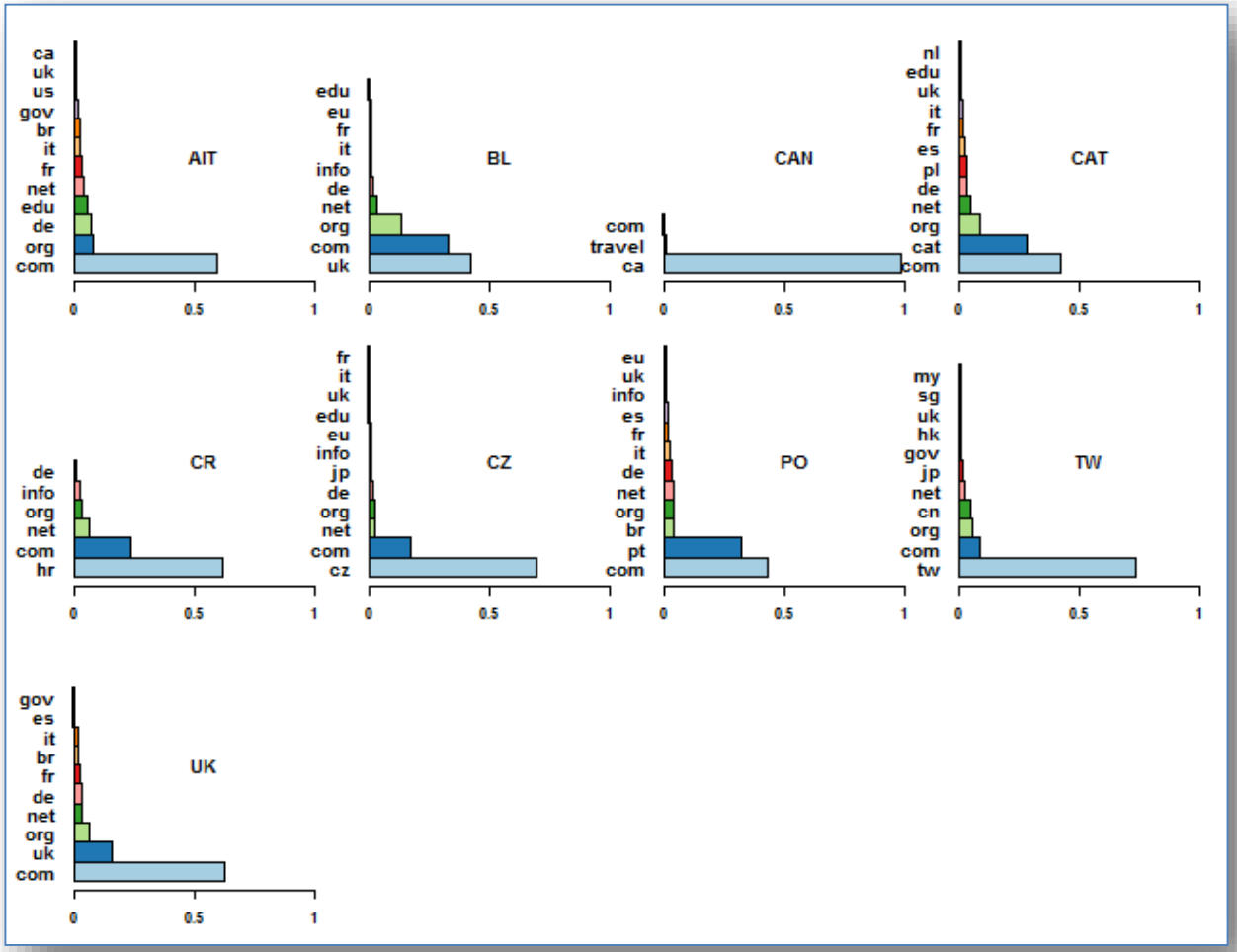
Entire Sample

TLD Distribution per Archive



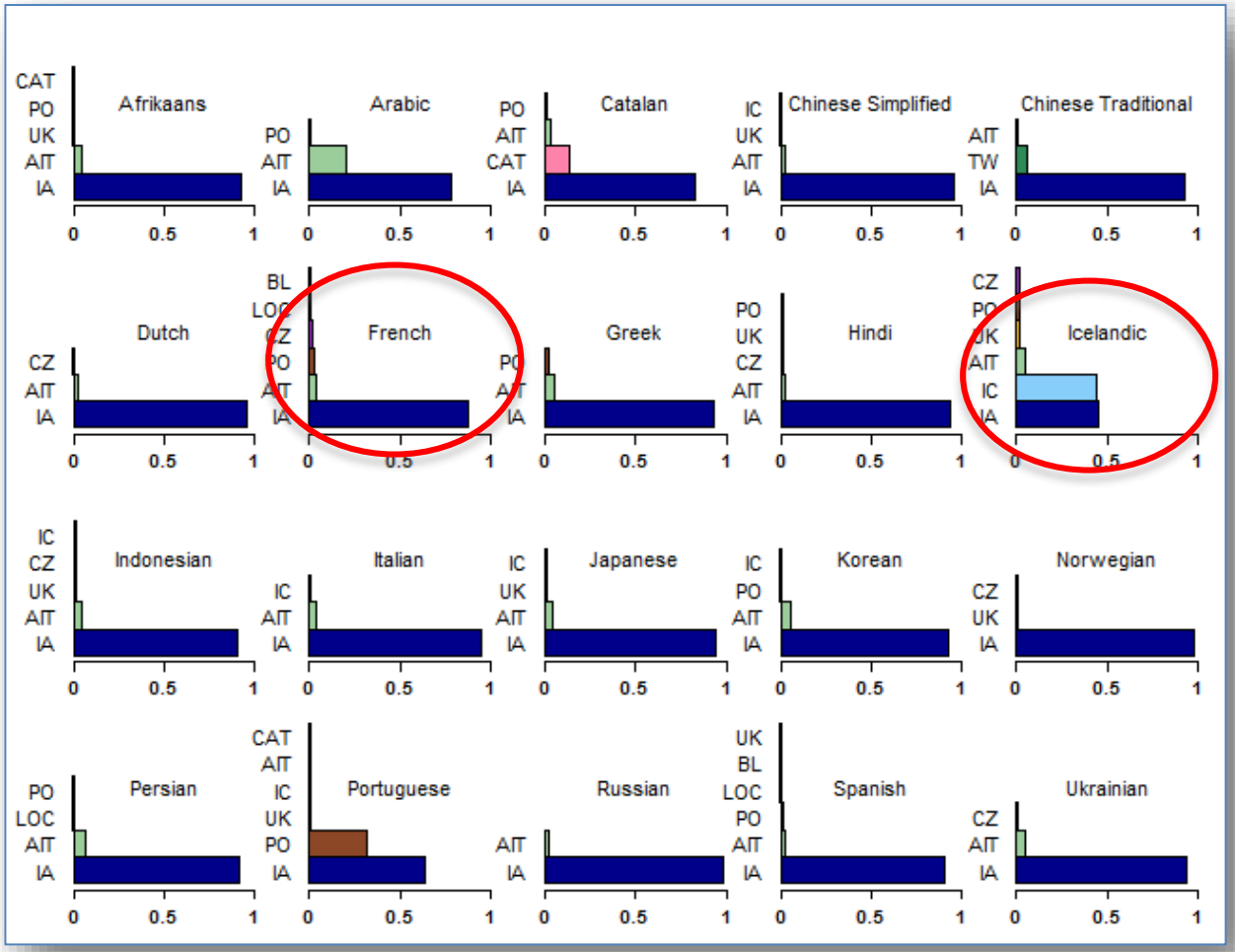
DMOZ:TLD Sample

TLD Distribution per Archive



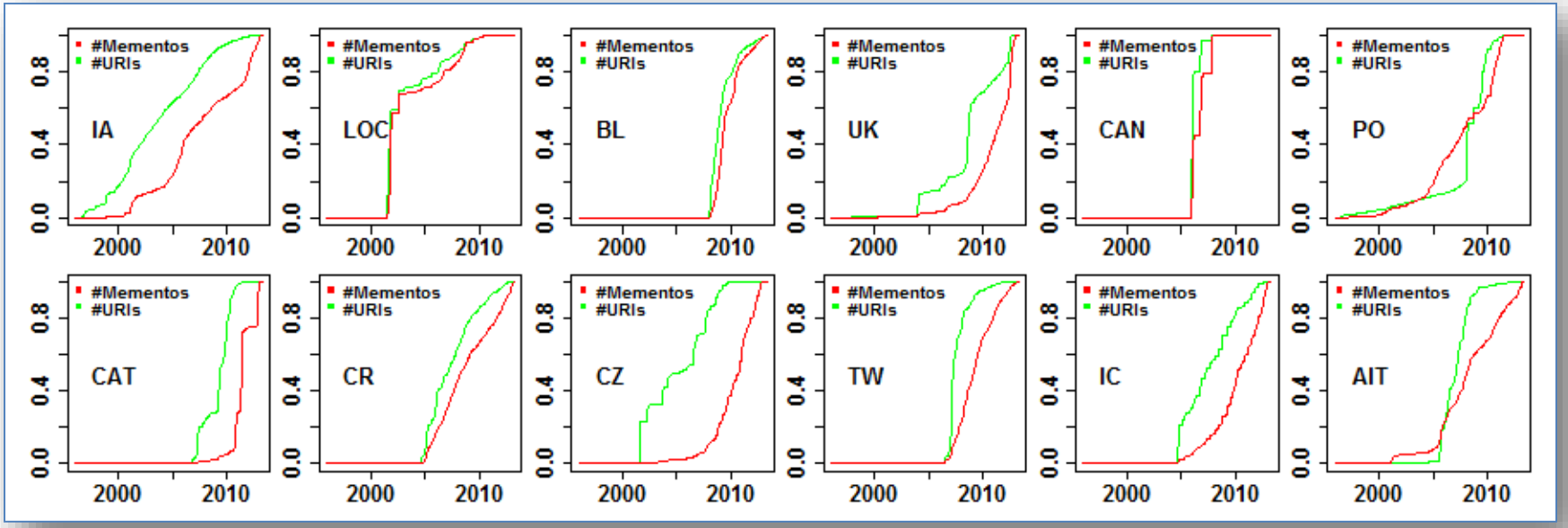
Web Archives Full Text Sample

Language Coverage per Archive



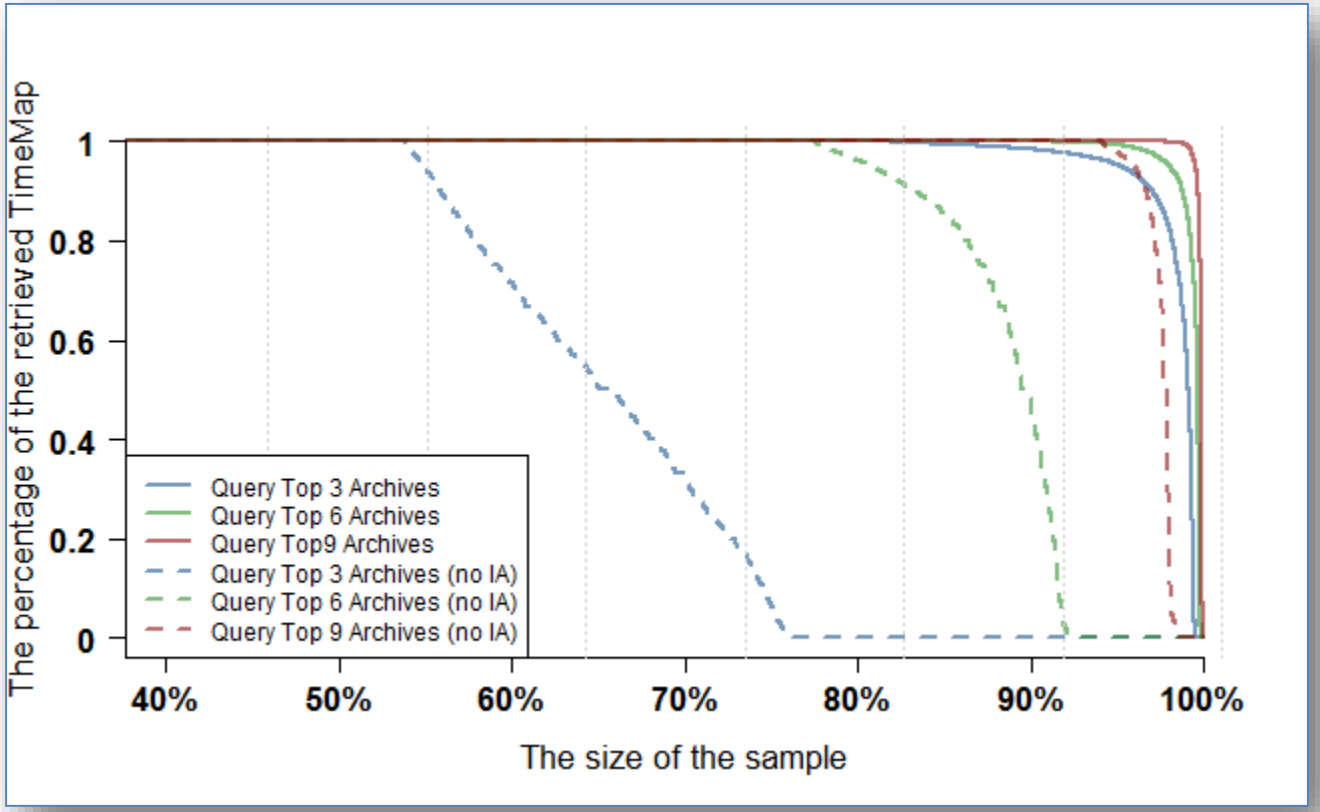
DMOZ Sample

Archive Growth Rate



Entire Sample

Query Routing Evaluation



Study Results

- Introduced sampling to profile web archives using available infrastructure, no privileged access

- Coverage:
  - Internet Archive provides broad coverage
  - National archives have good coverage for their domains
  - Surprising coverage by certain archives

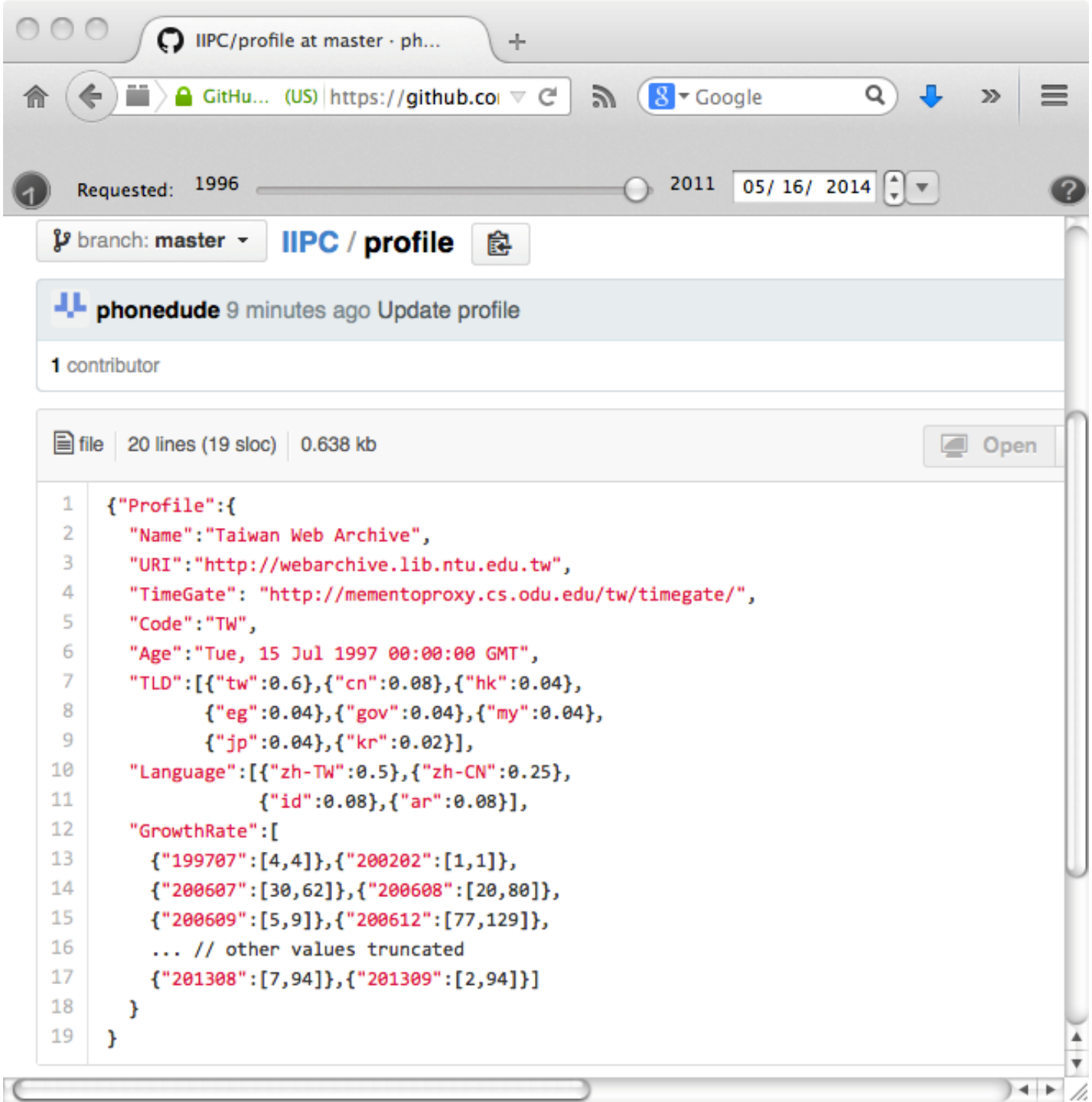
- Query Routing:
  - In 84% of the cases, all existing Mementos for a TLD can be found by using IA and two additional top archives for a TLD
  - In 55% of the cases, all existing Mementos for a TLD can be found by using the top 3 archives for a TLD, excluding IA

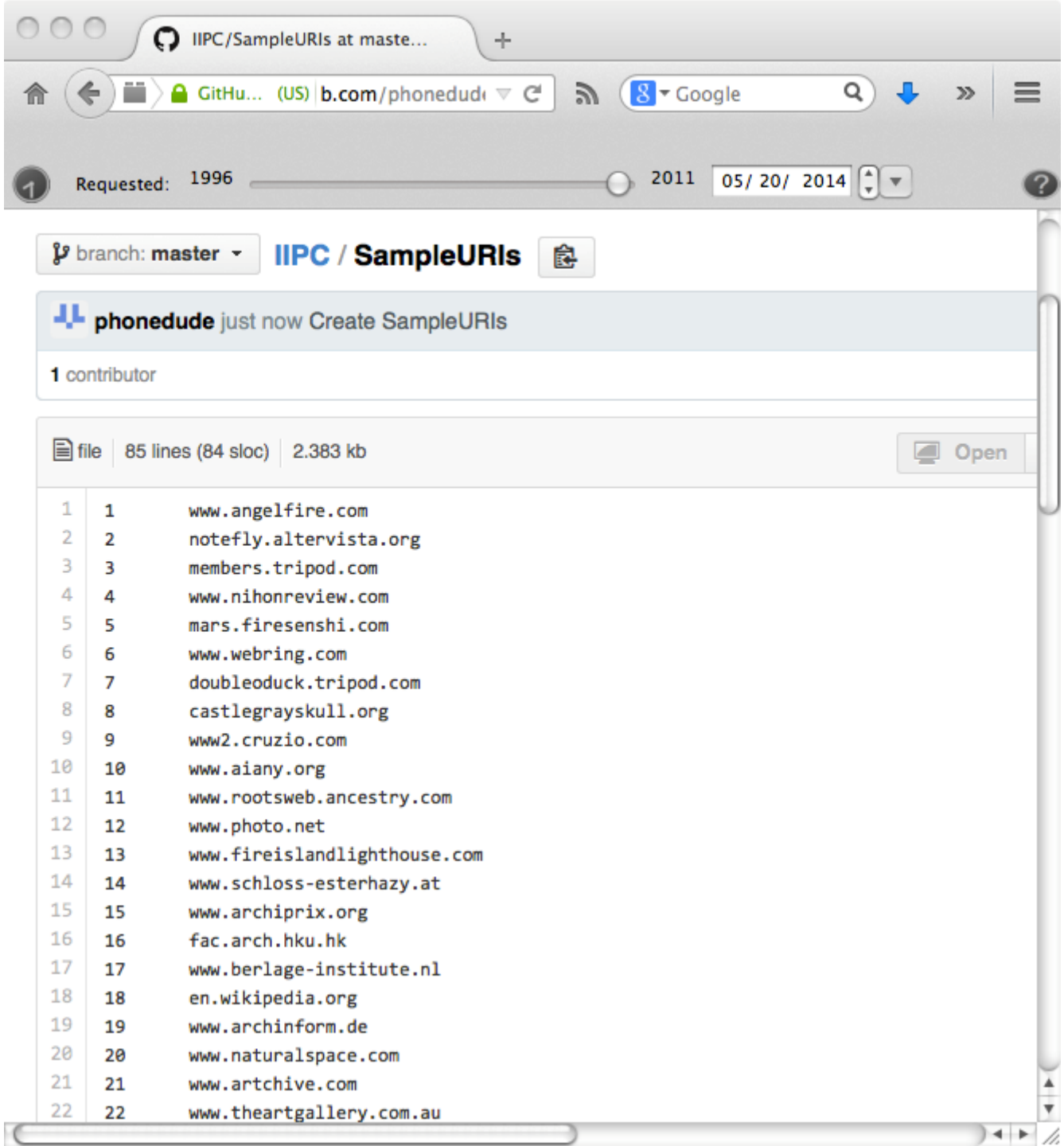
Next Steps With the IIPC

- Finding the right granularity
  - too fine: http://www.bnf.fr/fr/evenements_et_culture/a.passe_bnf.html
  - too coarse: .fr
  - just right?: bnf.fr, www.bnf.fr, gallica.bnf.fr, www.bnf.fr/fr/
- Generating profiles
  - what are desirable / representative sample sets: domains, languages, regions, etc. -- what's missing?
  - local CDX analysis tools (can help with cold start problem)
- Profile format
  - community input (yet another metadata format)
  - github (or other tools) for exchange & integration

A Possible Serialization

```
{"Profile":{
  "Name":"Taiwan Web Archive",
  "URI":"http://webarchive.lib.ntu.edu.tw",
  "TimeGate":
  "http://mementoproxy.cs.odu.edu/tw/timegate/",
  "Code":"TW",
  "Age":"Tue, 15 Jul 1997 00:00:00 GMT",
  "TLD":[{"tw":0.6},{"cn":0.08},{"hk":0.04},
         {"eg":0.04},{"gov":0.04},{"my":0.04},
         {"jp":0.04},{"kr":0.02}],
  "Language":[{"zh-TW":0.5},{"zh-CN":0.25},
              {"id":0.08},{"ar":0.08}],
  "GrowthRate":[
    {"199707":[4,4]},{"200202":[1,1]},
    {"200607":[30,62]},{"200608":[20,80]},
    {"200609":[5,9]},{"200612":[77,129]},
    ... // other values truncated
    {"201308":[7,94]},{"201309":[2,94]}]
  }
}
```


IIPC/profile at master · ph...
GitHu... (US) https://github.co
Google
Requested: 1996 2011 05/ 16/ 2014
branch: master
IIPC / profile
phonedude 9 minutes ago Update profile
1 contributor
file 20 lines (19 sloc) 0.638 kb
Open
{"Profile":{
"Name":"Taiwan Web Archive",
"URI":"http://webarchive.lib.ntu.edu.tw",
"TimeGate": "http://mementoproxy.cs.odu.edu/tw/timegate/",
"Code":"TW",
"Age":"Tue, 15 Jul 1997 00:00:00 GMT",
"TLD":[{"tw":0.6},{"cn":0.08},{"hk":0.04},
{"eg":0.04},{"gov":0.04},{"my":0.04},
{"jp":0.04},{"kr":0.02}],
"Language":[{"zh-TW":0.5},{"zh-CN":0.25},
{"id":0.08},{"ar":0.08}],
"GrowthRate":[
{"199707":[4,4]},{"200202":[1,1]},
{"200607":[30,62]},{"200608":[20,80]},
{"200609":[5,9]},{"200612":[77,129]},
... // other values truncated
{"201308":[7,94]},{"201309":[2,94]}]
}
}


IIPC/SampleURIs at maste...
GitHu... (US) b.com/phonedud
Google
Requested: 1996 2011 05/ 20/ 2014
branch: master
IIPC / SampleURIs
phonedude just now Create SampleURIs
1 contributor
file 85 lines (84 sloc) 2.383 kb
Open
1 1 www.angelfire.com
2 2 notefly.altervista.org
3 3 members.tripod.com
4 4 www.nihonreview.com
5 5 mars.firesenshi.com
6 6 www.webring.com
7 7 doubleoduck.tripod.com
8 8 castlegrayskull.org
9 9 www2.cruzio.com
10 10 www.aiany.org
11 11 www.rootsweb.ancestry.com
12 12 www.photo.net
13 13 www.fireislandlighthouse.com
14 14 www.schloss-esterhazy.at
15 15 www.archiprix.org
16 16 fac.arch.hku.hk
17 17 www.berlage-institute.nl
18 18 en.wikipedia.org
19 19 www.archinform.de
20 20 www.naturalspace.com
21 21 www.artchive.com
22 22 www.theartgallery.com.au

{Light, Dim, Dark} Archives

- Work to date has assumed light archives because our focus has been on sampling archives we don't control
- Applicable to a continuum of archives:
  - download/fork and run "dark-sample.py"
  - it accesses sample URIs from IIPC github
  - issues URI lookups to local archive
  - write/update your archive profile in IIPC github with machine readable IP restrictions
  - all profiles -- light/dim/dark -- now available to Memento aggregators *and other IIPC analysis tools*

Profiles = Easy Discovery, Sharing

http://netpreserve.org/aggr/timemap/link/1/http://www.bnf.fr/